Assignment 09: Data Scraping

Aubrey Knier

Total points:

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_09_Data_Scraping.Rmd") prior to submission.

Set up

1. Set up your session:

- Check your working directory
- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Set your ggplot theme

```
#1
getwd()
```

```
## [1] "/Users/aubreyknier/Desktop/Spring 2022/ENV872_EDA/Environmental_Data_Analytics_2022/Assignments"
```

```
library(tidyverse)
library(rvest)
library(lubridate)

mytheme <- theme_light() +
  theme(axis.text = element_text(color = "darkorchid4"),
        legend.position = "right")
theme_set(mytheme)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2020 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Scroll down and select the LWSP link next to Durham Municipality.
- Note the web address: https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
website <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020")
website
```

```
## {html_document}
## <html xmlns="http://www.w3.org/1999/xhtml" lang="en" xml:lang="en">
## [1] <head>\n<title>DWR :: Local Water Supply Planning</title>\n<meta http-equ ...
## [2] <body id="plan">\r\n<!--<div id="division-header">\r\n<a name="top" href= ...
```

3. The data we want to collect are listed below:

- From the "1. System Information" section:
- Water system name
- PSWID
- Ownership
- From the "3. Water Supply Sources" section:
- MAX Daily Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to three separate variables.

> HINT: The first value should be "Durham", the second "03-32-010", the third "Municipality", and the last should be a vector of 12 numeric values, with the first value being 36.0100.

```
#3
water.system.name <- website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
water.system.name
```

```
## [1] "Durham"
```

```
pswid <- website %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
pswid
```

```
## [1] "03-32-010"
```

```
ownership <- website %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
ownership
```

```
## [1] "Municipality"
```

```
max.withdrawals.mgd <- website %>%
  html_nodes("th~ td+ td") %>%
  html_text()
max.withdrawals.mgd
```

```
##  [1] "36.0100" "36.9800" "41.6900" "32.0500" "40.6100" "40.5600" "37.2900"
##  [8] "43.6300" "33.3200" "32.3700" "41.9300" "28.0600"
```

4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

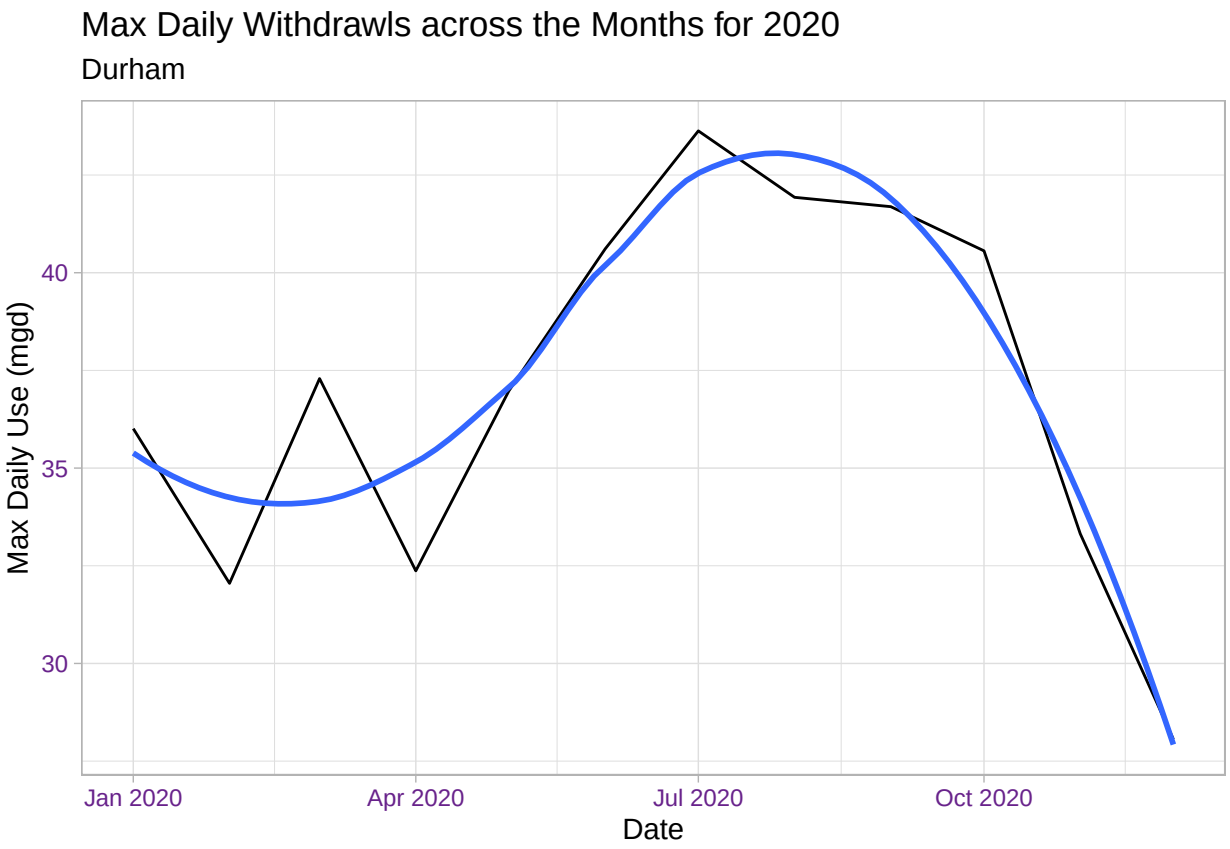
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc...

5. Plot the max daily withdrawals across the months for 2020

```
#4
watersupply.df <- data.frame("Month" = c("Jan", "May", "Sept", "Feb",
                                         "Jun", "Oct", "Mar", "Jul",
                                         "Nov", "Apr", "Aug", "Dec"),
                             "Year" = 2020,
                             "Water_System_Name" = water.system.name,
                             "PSWID" = pswid,
                             "Ownership" = ownership,
                             "Max_Daily_Use_per_Month" = max.withdrawals.mgd
                              )
watersupply.df <- watersupply.df %>%
                  mutate(Date=my(paste(Month, "-", Year)))


#5
ggplot(watersupply.df, aes(x=Date,y=as.numeric(Max_Daily_Use_per_Month), group=1)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Max Daily Withdrawls across the Months for 2020"),
       subtitle = water.system.name,
       y="Max Daily Use (mgd)",
       x="Date")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

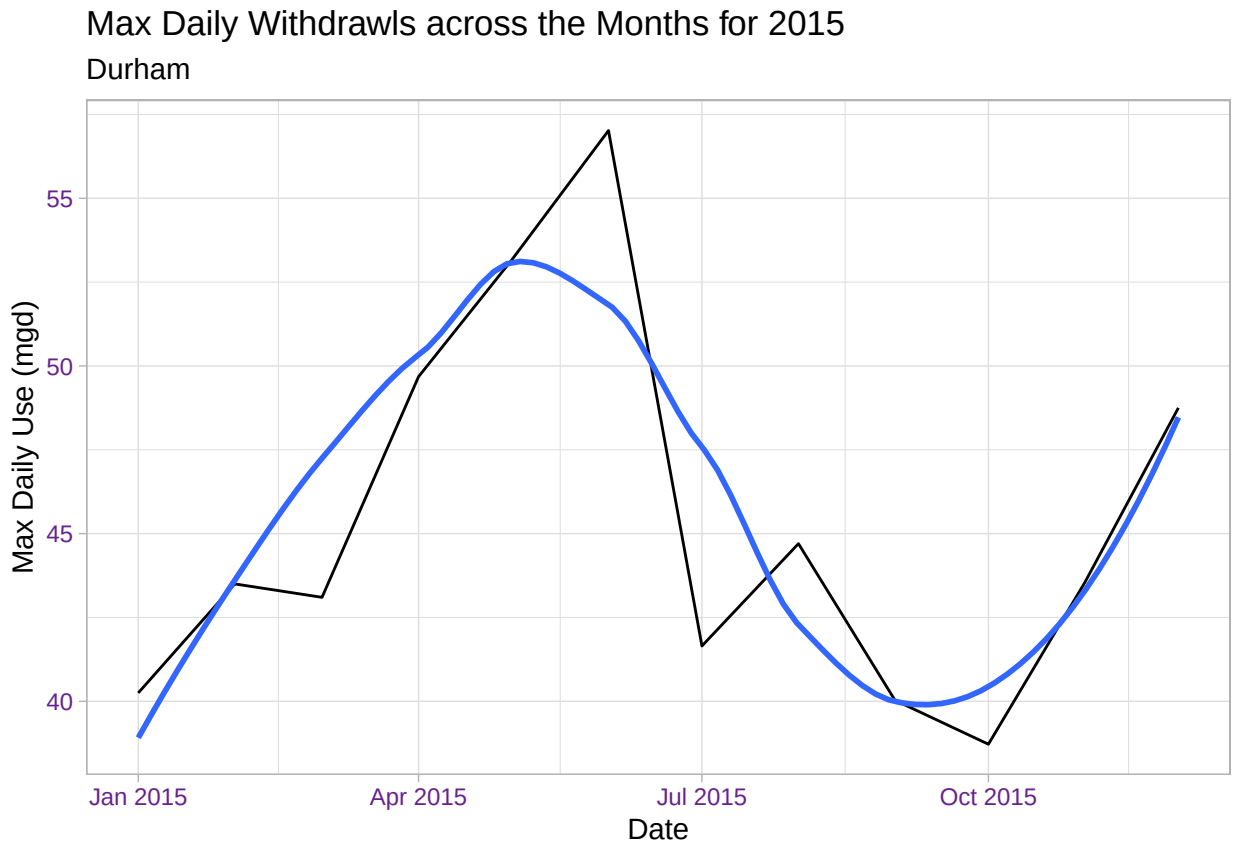
```
#6.
scrape.it <- function(the.year, pswid){

  the.website <- read_html(paste0(
    'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',pswid,'&year=',
    the.year))

  water.system.name.tag <- "div+ table tr:nth-child(1) td:nth-child(2)"
  pswid.tag <- "td tr:nth-child(1) td:nth-child(5)"
  ownership.tag <- "div+ table tr:nth-child(2) td:nth-child(4)"
  max.withdrawals.mgd.tag <- "th~ td+ td"

  water.system.name <- the.website %>% html_nodes(water.system.name.tag) %>% html_text()
  pswid <- the.website %>% html_nodes(pswid.tag) %>% html_text()
  ownership <- the.website %>% html_nodes(ownership.tag) %>% html_text()
  max.withdrawals.mgd <- the.website %>% html_nodes(max.withdrawals.mgd.tag) %>% html_text()

  watersupply.df <- data.frame("Month" = c("Jan", "May", "Sept", "Feb",
                                           "Jun", "Oct", "Mar", "Jul",
```

```
                                          "Nov", "Apr", "Aug", "Dec"),
                              "Year" = the.year,
                              "Water_System_Name" = water.system.name,
                              "PSWID" = pswid,
                              "Ownership" = ownership,
                              "Max_Daily_Use_per_Month" = max.withdrawals.mgd
                               )
watersupply.df <- watersupply.df %>%
                  mutate(Date=my(paste(Month, "-", Year)))



return(watersupply.df)

}
```

7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
new.df <- scrape.it(2015, '03-32-010')
view(new.df)

ggplot(new.df, aes(x=Date,y=as.numeric(Max_Daily_Use_per_Month, group=1))) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Max Daily Withdrawls across the Months for 2015"),
       subtitle = water.system.name,
       y="Max Daily Use (mgd)",
       x="Date")
```

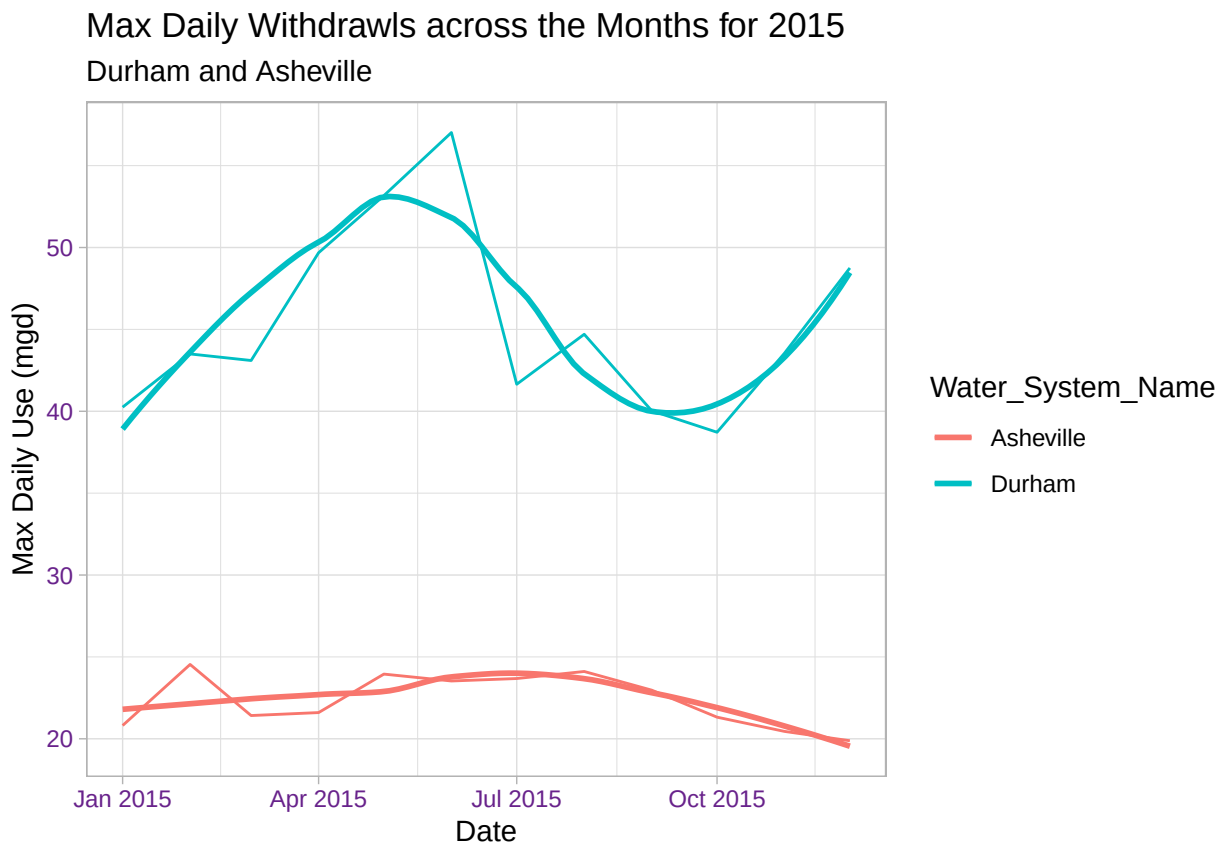
```
## `geom_smooth()` using formula 'y ~ x'
```



8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

```
#8
asheville.df <- scrape.it(2015, '01-11-010')
dur.ash.df <- rbind(new.df, asheville.df)

ggplot(dur.ash.df, aes(x=Date, y=as.numeric(Max_Daily_Use_per_Month), group=Water_System_Name,
                       color=Water_System_Name)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Max Daily Withdrawls across the Months for 2015"),
       subtitle = "Durham and Asheville",
       y="Max Daily Use (mgd)",
       x="Date")
```

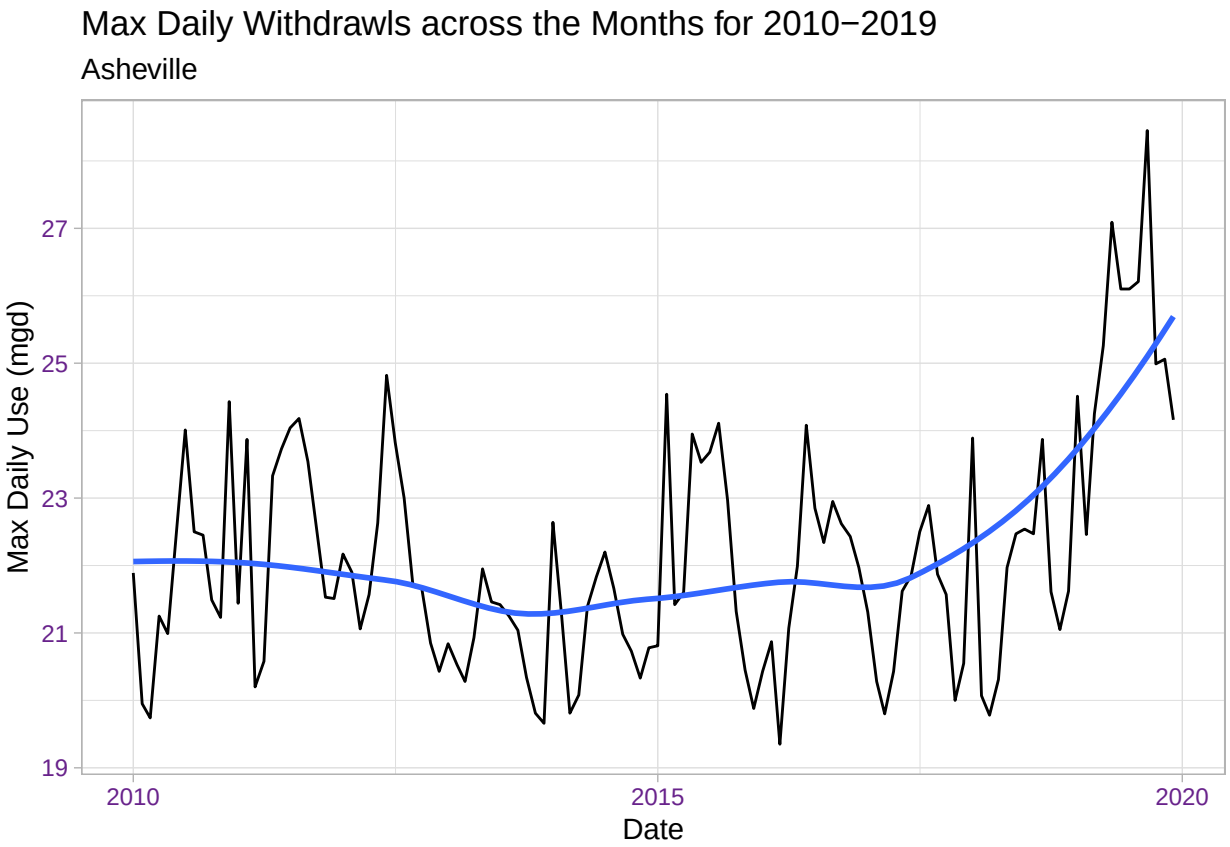
```
## `geom_smooth()` using formula 'y ~ x'
```



9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
the.years = rep(2010:2019)
pswid = '01-11-010'

ash.1019.dfs <- lapply(X = the.years,
                  FUN = scrape.it,
                  pswid=pswid)

ash.1019.df <- bind_rows(ash.1019.dfs)

ggplot(ash.1019.df,aes(x=Date,y=as.numeric(Max_Daily_Use_per_Month), group=1)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Max Daily Withdrawls across the Months for 2010-2019"),
       subtitle = "Asheville",
       y="Max Daily Use (mgd)",
       x="Date")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? Yes, maximum daily water use in Asheville has increased over the years 2010-2019.